オーバーサンプリングを用いた括弧表現分類における データ生成モデルの精度についての考察

情報科学科　村田 浩章　　指導教員：山村 毅

1　はじめに

日本語で記述される文章には，丸括弧や鉤括弧などの括弧表現が多く使われる．文章中の括弧表現には様々な用法が存在し，同じ丸括弧や鍵括弧でも複数の用法で使用される．用法の違いは文法的な役割の違いでもあるため，文章に対して構文解析等の機械処理を行う際にも，括弧表現の扱いを用法ごとに変える必要がある．

先行研究では 6 つの括弧表現カテゴリを提案し，有効な素性を抽出し，最大エントロピーモデルを用いて分類を行った．

本研究では先行研究で問題にされていた不均衡なデータをオーバーサンプリングを用いて解消し，またオーバーサンプリングにおけるデータ生成モデルの精度が分類結果にどのような影響を与えるのか考察する．

2　先行研究 [1]

2.1　括弧表現の用法

本文に対して括弧記号が必要か不必要か，括弧表現が必要か不必要か，またその処理方法の違いに基づいて補足情報，引用，独立文，小見出し，補完，記号表現の 6 つの分類カテゴリを設けた．

2.2　括弧表現の処理

2005 年の毎日新聞の 5 日分の記事から括弧表現を 7809 個抽出し，2.1 節で述べた分類カテゴリを人手でラベル付けしたものを用いて，分類実験を行った．各カテゴリのデータ数を表 1 に示す (表中の補足は補足情報，独立は独立文，小見は小見出し，記号は記号表現を表す).

表 1　各カテゴリのデータ数

ラベル	補足	引用	独立	小見	補完	記号
数	3382	2915	438	640	103	331

括弧表現の文における位置や括弧記号の形などに着目した素性を 6 個抽出し，これを特徴量として括弧表現の分類を行い，分類精度を調べた．分類器には最大エントロピーモデルを用い，分類精度の計算には 10 分割交差検定を用いた．先行研究の結果を表 2 に示す．

表 2　先行研究

	補足	引用	独立	小見	補完	記号	平均
再現率	.960	.983	.532	.675	.000	.453	.601
精度	.882	.978	.766	.614	.000	.785	.671
F 値	.919	.981	.628	.643	.000	.575	.624

実験の結果，全体として 88.7% という精度を得た．だが補完カテゴリの精度は 0% と極めて悪かった．これは実験データにおいて補完のデータが少なく，データに大きな偏りがあるためだと考えられる．この偏りを解消するために，少数データに対してオーバーサンプリングをする必要がある．

3　実験と結果

先行研究と同じデータ，素性を用いて，データの少ないカテゴリについてオーバーサンプリングをした上で分類実験を行った．データ数の最も多い補足カテゴリの 3382 個になるまで，各カテゴリにおいてデータの生成確率を計算し，それに合わせてオーバーサンプリングをした．3 つのデータ生成モデルを考え精度を比較した．

1. 括弧及び，全ての素性に独立に生成した場合．
2. 括弧記号を先に生成し，その他の素性は括弧記号にのみ依存して生成した場合．
3. 括弧及び，全ての素性に依存して生成した場合．

これらについて，10 分割交差検定を用いて実験を行った結果を表 3，表 4，表 5 に示す．また，全体の正解率はそれぞれ 69.7%，69.6%，69.9% であった．

表 3　括弧及び全ての素性に独立に生成

	補足	引用	独立	小見	補完	記号	平均
再現率	.544	.983	.642	.231	.942	.625	.661
精度	.970	.982	.643	.733	.055	.342	.621
F 値	.697	.983	.642	.352	.105	.442	.537

表 4　括弧記号にのみ依存して生成

	補足	引用	独立	小見	補完	記号	平均
再現率	.544	.981	.642	.231	.942	.625	.661
精度	.967	.982	.643	.733	.055	.342	.620
F 値	.696	.982	.642	.352	.105	.442	.536

表 5　括弧及び全ての素性に依存して生成

	補足	引用	独立	小見	補完	記号	平均
再現率	.554	.981	.642	.231	.942	.604	.659
精度	.964	.982	.643	.733	.055	.355	.622
F 値	.703	.982	.642	.352	.105	.447	.539

4　まとめ

本研究では，データ数の少ないカテゴリについてオーバーサンプリングをし，最大エントロピーモデルを用いて分類を行った．

表 3〜表 5 から分かるように，今回の実験では，データ生成モデルの精度を上げても結果はほとんど変わらなかったといえる．

またオーバーサンプリングによってデータの偏りを解消したものの，補完カテゴリについての精度はほとんど上がらなかった．

先行研究と比較して，どのオーバーサンプリングの方法でも，特に補足，小見出しカテゴリの F 値が大幅に下がっていることから，選択した 6 個の特徴量はそもそも適切ではなかったといえる．一方で引用カテゴリにおいては先行研究，本研究どちらも F 値が高く，引用カテゴリを分類するのには適した特徴であるといえる．

今後はオーバーサンプリングすることを前提に，改めて特徴量を決定していく必要がある．

参考文献

[1] 村田 , 落合 , 山村 ”最大エントロピーモデルを用いた括弧表現の分類”, 電気・電子・情報関係学会東海支部連合大会講演論文集 , L4-2 , 2014